

# Estatística Inferencial

Prof. Wagner Hugo Bonat

Departamento de Estatística  
Universidade Federal do Paraná





# **Importância, conceitos, elementos e aplicações**

- ▶ **Registros (do Estado):**
  - ▶ **estatísticas oficiais,**
  - ▶ **inventários.**
- ▶ Método científico e ciência estatística:
  - ▶ probabilidades e modelos,
  - ▶ modelagem e incerteza,
  - ▶ apoio à decisão.
- ▶ Está em toda parte:
  - ▶ *on the fly*,
  - ▶ *Big Data*,
  - ▶ *Data Science*.



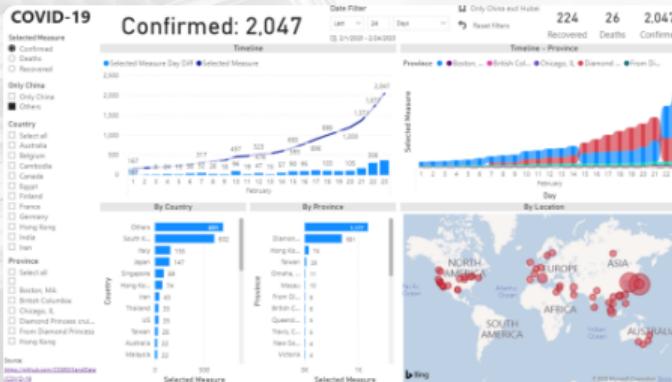
# Estatística

- ▶ Registros (do Estado):
  - ▶ estatísticas oficiais,
  - ▶ inventários.
- ▶ **Método científico e ciência estatística:**
  - ▶ **probabilidades e modelos,**
  - ▶ **modelagem e incerteza,**
  - ▶ **apoio à decisão.**
- ▶ Está em toda parte:
  - ▶ *on the fly,*
  - ▶ *Big Data,*
  - ▶ *Data Science.*



# Estatística

- ▶ Registros (do Estado):
    - ▶ estatísticas oficiais,
    - ▶ inventários.
  - ▶ Método científico e ciência estatística:
    - ▶ probabilidades e modelos,
    - ▶ modelagem e incerteza,
    - ▶ apoio à decisão.
  - ▶ **Está em toda parte:**
    - ▶ *on the fly*,
    - ▶ *Big Data*,
    - ▶ *Data Science*.



# O que é estatística?

- ▶ Estatística é um conjunto de técnicas para, sistematicamente:
  - ▶ Planejar a coleta de dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento;
  - ▶ Descrever, analisar e interpretar dados;
  - ▶ Extrair informações para subsidiar decisões;
  - ▶ Avaliar evidências empíricas sob hipóteses de interesse.
- ▶ Exemplos de aplicações:
  - ▶ Opinião da população brasileira sobre uma proposta do governo;
  - ▶ Avaliar a efetividade de uma nova droga para o tratamento de uma doença;
  - ▶ Entender os hábitos de compra dos clientes de uma loja virtual;
  - ▶ Recomendar produtos de forma personalizada;
  - ▶ Comparar a produtividade de uma cultivar sob diferentes formas de cultivo, adubação, etc.

# Conceitos fundamentais

- ▶ **População:** Conjunto de todos os elementos sob investigação.
- ▶ **Amostra:** Subconjunto da população.
- ▶ **Variável** de interesse: característica a ser observada em cada indivíduo da amostra.



Figura



Figura 2. Extraído de Pixabay.

# Exemplos em detalhes

- ▶ Opinião da população brasileira sobre uma proposta do governo.
  - ▶ **População:** Todos os habitantes do Brasil? outras opções?
  - ▶ **Amostra:** Algum subconjunto da população. Qualquer um será? Como selecionar?
  - ▶ **Variável de interesse:** Como medir isso? Concorda? sim(1) ou não(0).
- ▶ Avaliar a efetividade de uma nova droga para tratamento de uma forma de câncer.
  - ▶ **População:** Todos os seres humanos? Apenas os já doentes? Como levar em conta questões de raça, culturas, etc ...
  - ▶ **Amostra:** E agora?
  - ▶ **Variável de interesse:** Curou ou não curou? Será que isso é possível?
- ▶ Entender os hábitos de compra dos clientes de uma loja virtual.
  - ▶ **População:** Todos os clientes da loja virtual.
  - ▶ **Amostra:** Preciso de amostra?
  - ▶ **Variável de interesse:** E agora? Como caracterizar hábito de compra?

# Temas da estatística

- ▶ Estatística descritiva ou exploratória:
  - ▶ Consistência dos dados e interpretações iniciais.
  - ▶ Visualização dos dados e relações entre variáveis.
- ▶ Probabilidade:
  - ▶ Fornece ferramentas para lidar/quantificar incerteza.
- ▶ **Inferência estatística:**
  - ▶ **Estimação de quantidades desconhecidas.**
  - ▶ **Formular e testar hipóteses.**
  - ▶ **Extrapolar para a população resultados obtidos na amostra.**



Figura 3. Extraído de pexels.com

# Modelagem estatística

# Tipos de fenômenos

## Fenômenos determinísticos

Dizemos que um fenômeno é determinístico quando repetido inúmeras vezes, **em condições semelhantes**, conduz a resultados **essencialmente** idênticos. Ex.:

- ▶ Aceleração da gravidade.
- ▶ Algumas leis da Física (mecânica clássica) e da Química.

## Fenômenos aleatórios

Os fenômenos **repetidos sob as mesmas condições** e que geram resultados diferentes são chamados de fenômenos aleatórios. Ex.:

- ▶ Lançamento de uma moeda, dado ou similar.
- ▶ Resultado de um evento esportivo.
- ▶ Condições climáticas do próximo domingo.

# Variável aleatória

Uma **variável aleatória** é uma variável que assume valores numéricos associados aos resultados de um experimento aleatório, onde um (e apenas um) valor numérico é associado com cada ponto do espaço amostral.

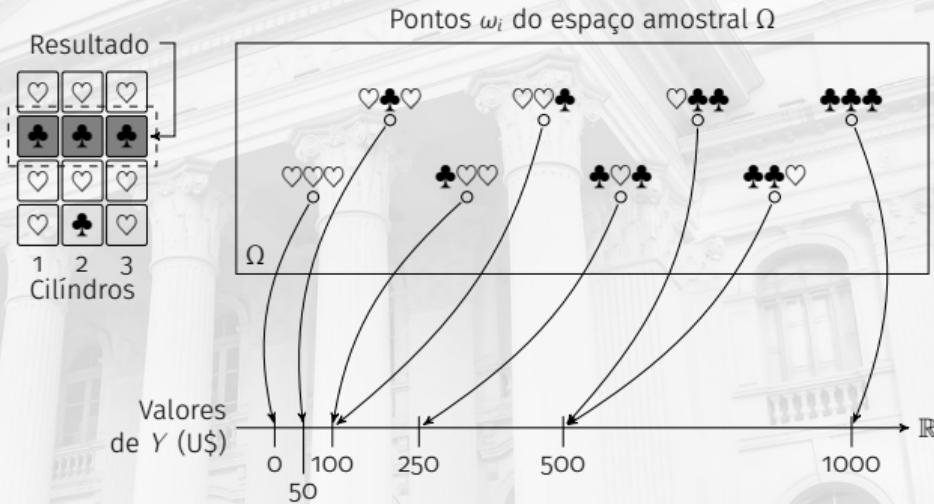


Figura 4. Exemplo de definição de variável aleatória.

# Motivação sobre os modelos probabilísticos

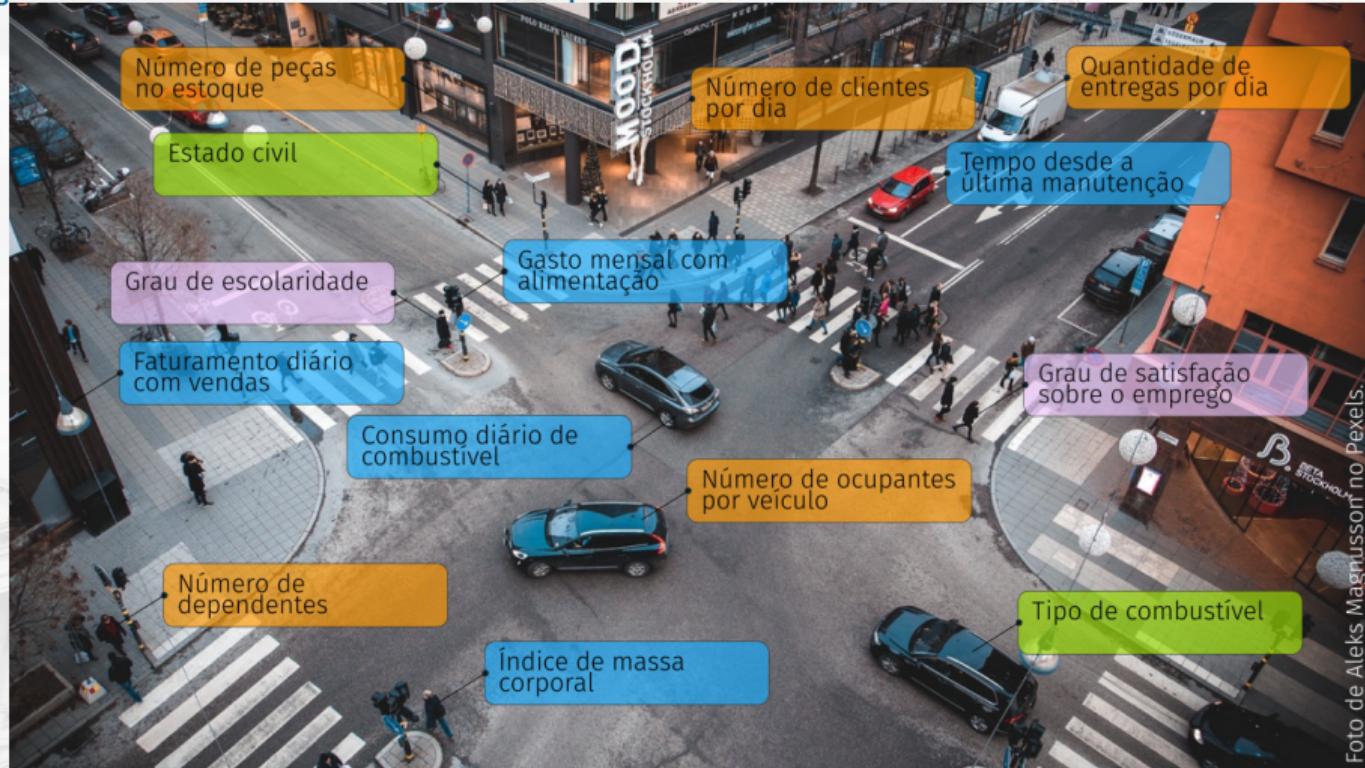


Figura 5. Variáveis aleatórias de uma típica cena cotidiana.

# Tipos de v.a's

## ► Espaço amostral

- ▶ Reta real,  $\Omega = \mathbb{R}$ .
- ▶ Estritamente positivos,  $\Omega = \mathbb{R}_+$ .
- ▶ Positivos com zeros,  $\Omega = \mathbb{R}_0 = [0, \infty)$ .
- ▶ Proporções ou índices (limitadas)  $\Omega = (0, 1)$ .
- ▶ Direções,  $\Omega = [0, 2\pi)$ .
- ▶ Contagem,  $\Omega = \mathbb{N}_0 = \{0, 1, 2, \dots\}$ .
- ▶ Contagem limitada,  $\Omega = \{0, 1, 2, \dots, m\}$ .

## ► Tipos de espaço amostral

▶ **Espaço amostral Discreto:** Contém apenas um número finito ou contável de elementos.

▶ **Espaço amostral Contínuo:** Contém um número infinito e não contável de elementos.

## ► Tipos de variáveis aleatórias

- ▶ Variável aleatória é **discreta** se seu espaço amostral é discreto.
- ▶ Variável aleatória é **contínua** se seu espaço amostral é contínuo.

# Distribuição de probabilidades

## ► Função de probabilidade (fp)

$$P(Y = y_i; \theta) = p(y_i) = p_i, \quad i = 1, 2, \dots$$

## ► Propriedades

- ▶  $0 \leq p(y_i) \leq 1, \quad \forall i = 1, 2, \dots$ .
- ▶  $\sum_i p(y_i) = 1.$

► Vetor de parâmetros  $\theta \in \Theta$ .

## ► Densidade de probabilidade (fdp)

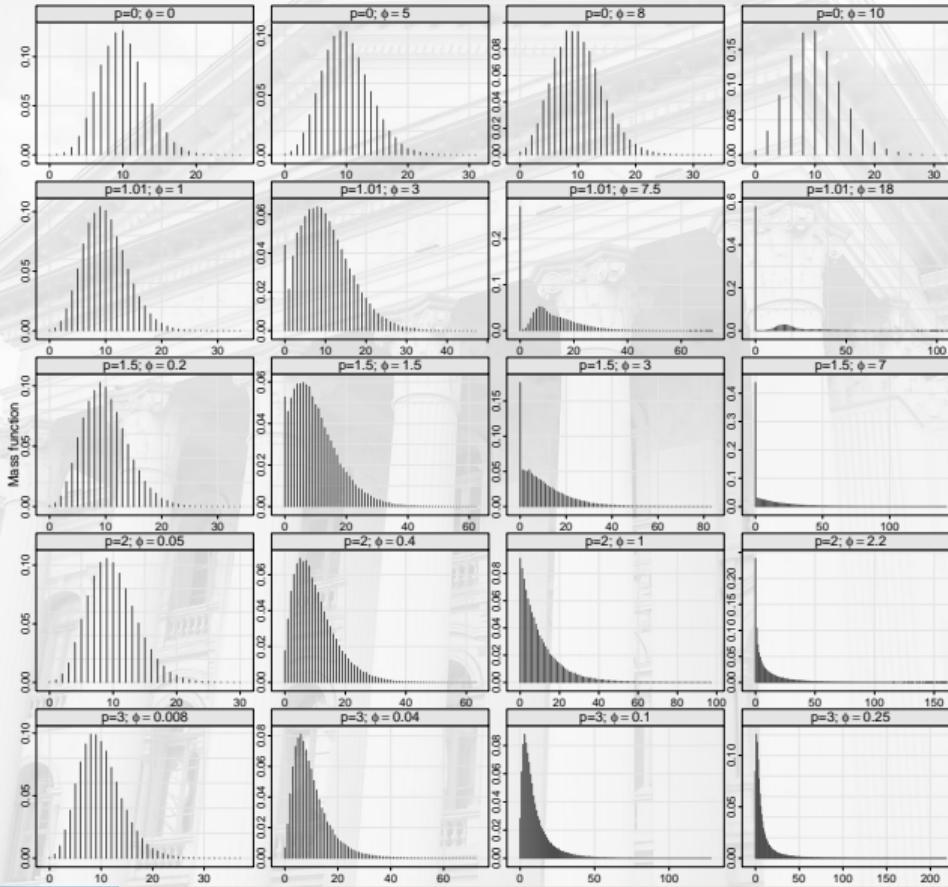
$$P(a < Y < b; \theta) = \int_a^b f(y; \theta) dy$$

## ► Propriedades:

- ▶  $f(y; \theta) \geq 0.$
- ▶  $\int_{-\infty}^{+\infty} f(y; \theta) dy = 1.$

► Vetor de parâmetros  $\theta \in \Theta$ .

# Formatos da distribuição de probabilidades



# Componentes de um modelo probabilístico

## Variável aleatória

- ▶ **Variável aleatória:** resultado numérico da observação de um fenômeno aleatório.
- ▶ **Suporte:** conjunto de valores que a variável aleatória  $Y$  pode assumir.
- ▶ Conforme o suporte, as variáveis aleatórias são:
  - ▶ Discretas.
  - ▶ Contínuas.

## Parâmetro

- ▶ **Parâmetro:** variável que é parte da distribuição de probabilidades.
- ▶ **Espaço paramétrico:** conjunto de valores válidos para o parâmetro da distribuição.
- ▶ Conforme o espaço paramétrico, os parâmetros são:
  - ▶ Discretos.
  - ▶ Contínuos.
- ▶ A distribuição pode ter qualquer quantidade de parâmetros, até mesmo nenhum.



# Motivação: Inferência estatística

# Inferência estatística

- ▶ **População** → distribuição de probabilidade.
- ▶ Intuição → Como que a v.a. deve se comportar na população.
- ▶ Variável → variável aleatória.
- ▶ Parâmetros da distribuição de probabilidade → **parâmetros populacionais**.
- ▶ Como obter a amostra?
- ▶ Como a partir da amostra estimar os parâmetros populacionais?

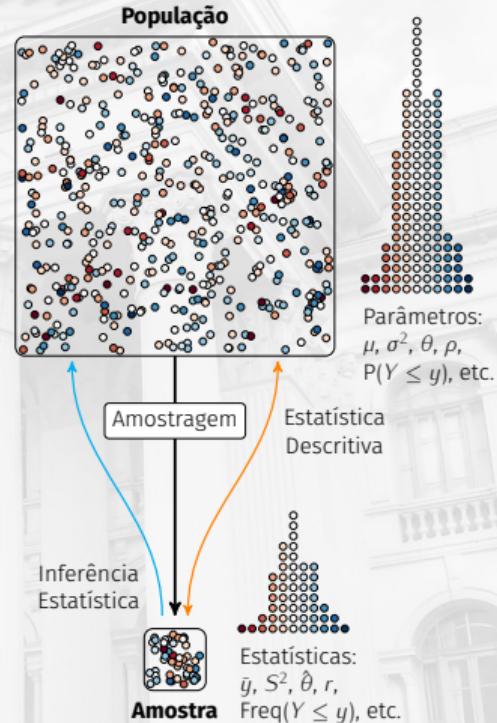


Figura 7. Processo de inferência estatística.

# Inferência estatística

- ▶ Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?

# Inferência estatística

- ▶ Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- ▶ Formalizando o problema:
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?

# Inferência estatística

- ▶ Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- ▶ Formalizando o problema:
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?
  - ▶  $Y$ : desenvolveu anticorpos. Opções SIM ou NÃO.
- ▶ Qual a distribuição de probabilidade adequada para esta v.a.?

# Inferência estatística

- ▶ Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- ▶ Formalizando o problema:
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?
  - ▶  $Y$ : desenvolveu anticorpos. Opções SIM ou NÃO.
- ▶ Qual a distribuição de probabilidade adequada para esta v.a.?
  - ▶ Bernoulli com função de probabilidade

$$P(Y = y) = p^y(1 - p)^{1-y}.$$

- ▶ Qual o parâmetro de interesse e o que ele significa?

# Inferência estatística

- ▶ Problema prático: Qual a proporção da população que desenvolveu anticorpos contra uma doença?
- ▶ Formalizando o problema:
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?
  - ▶  $Y$ : desenvolveu anticorpos. Opções SIM ou NÃO.
- ▶ Qual a distribuição de probabilidade adequada para esta v.a.?
  - ▶ Bernoulli com função de probabilidade

$$P(Y = y) = p^y(1 - p)^{1-y}.$$

- ▶ Qual o parâmetro de interesse e o que ele significa?
  - ▶  $p$ : proporção de pessoas que desenvolveram anticorpos.

- ▶ Como determinar o valor de  $p$ ?
  - ▶ Examinar todos os membros da população e verificar a proporção que desenvolveu anticorpos.
  - ▶ Examinar apenas alguns membros da população (amostra) e calcular a proporção que desenvolveu anticorpos.
- ▶ Problema: A proporção obtida na amostra não é a mesma obtida na população.
  - ▶ Incerteza associada ao valor da proporção devido a termos apenas uma amostra.
  - ▶ Como quantificar essa incerteza?
  - ▶ Como tomar uma decisão baseada apenas na amostra?
- ▶ Descrição probabilística da estatística de interesse → **Distribuição amostral.**

# Especificação do problema de Inferência

- ▶  $Y$ : desenvolveu anticorpos (v.a.).
- ▶ Especificação do modelo  $Y \sim \text{Ber}(p)$ .
- ▶ Parâmetro  $p$ .
- ▶ Informação sobre  $p$  através de uma amostra da população.
- ▶ Denotamos as amostras por  $y_1, \dots, y_n$ .
- ▶ Objetivos da inferência estatística:
  - ▶ Estimar  $p$  baseado apenas na amostra (valor pontual)! Quanto é  $p$  na população?
  - ▶ Informar o quanto preciso ou creditável é o valor estimado (intervalo de confiança).
  - ▶ Decidir sobre possíveis valores de  $p$  baseado apenas na amostra.
  - ▶ A proporção da população com anticorpos atingiu um patamar desejável?

# Especificação do problema de Inferência

- ▶ Suponha que coletamos uma amostra (aleatória) de tamanho  $n = 10$  e que  $y = 7$  pessoas apresentaram anticorpos.
- ▶ Qual valor você acha que o parâmetro  $p$  assume na população?
- ▶ Assumindo observações independentes, sabemos que a soma de v.a. Bernoulli é binomial com  $n = 10$  e um parâmetro  $p$  desconhecido.
- ▶ Podemos calcular a probabilidade de observar  $y = 7$  para um valor de  $p$ , por exemplo,  $p = 0.8$

$$P(Y = 7|n = 10, p = 0,80) = \binom{10}{7} 0,80^7(1 - 0,80)^{10-7} = 0,2013.$$

# Especificação do problema de Inferência

- ▶ Para qualquer outro valor de  $p$

$$P(Y = 7|n = 10, p) = \binom{10}{7} p^7 (1 - p)^{10-7},$$

variando  $p$  temos a **função de verossimilhança**

$$L(p) \equiv P(Y = 7|n = 10, p) = \binom{10}{7} p^7 (1 - p)^{10-7}.$$

- ▶ **Ideia:** Se  $p$  for um determinado valor, **qual a probabilidade** de observar o que eu realmente observei na amostra.

# Pensamento frequentista

- ▶ Se o experimento for repetido um número grande de vezes e a cada realização  $\hat{p}$  for obtido, o que aconteceria?
- ▶  $\hat{p}$  é uma variável aleatória.
- ▶ Se é variável aleatória, então tem distribuição de probabilidade que descreve o seu comportamento.
  - ▶ Qual é a sua distribuição?
  - ▶ Qual o seu valor esperado?
  - ▶ Qual a sua variância?

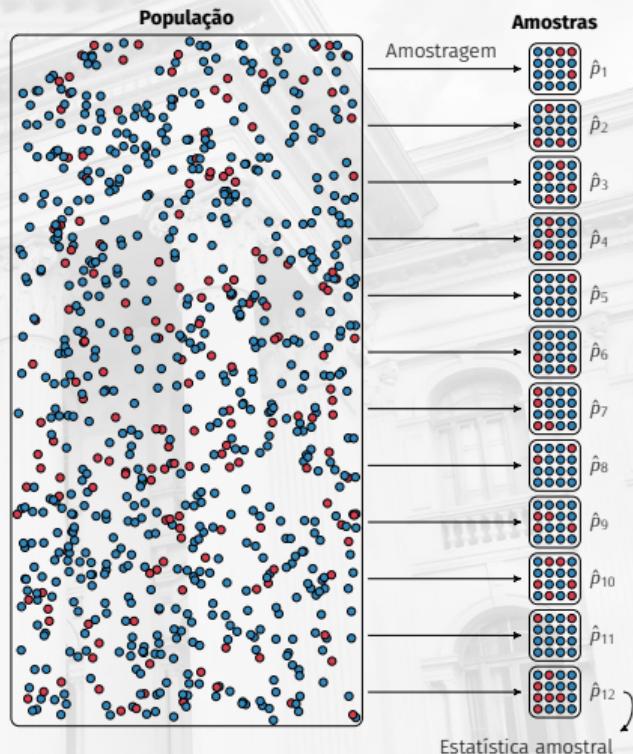


Figura 8. Ilustração da distribuição amostral.

# Ilustração computacional

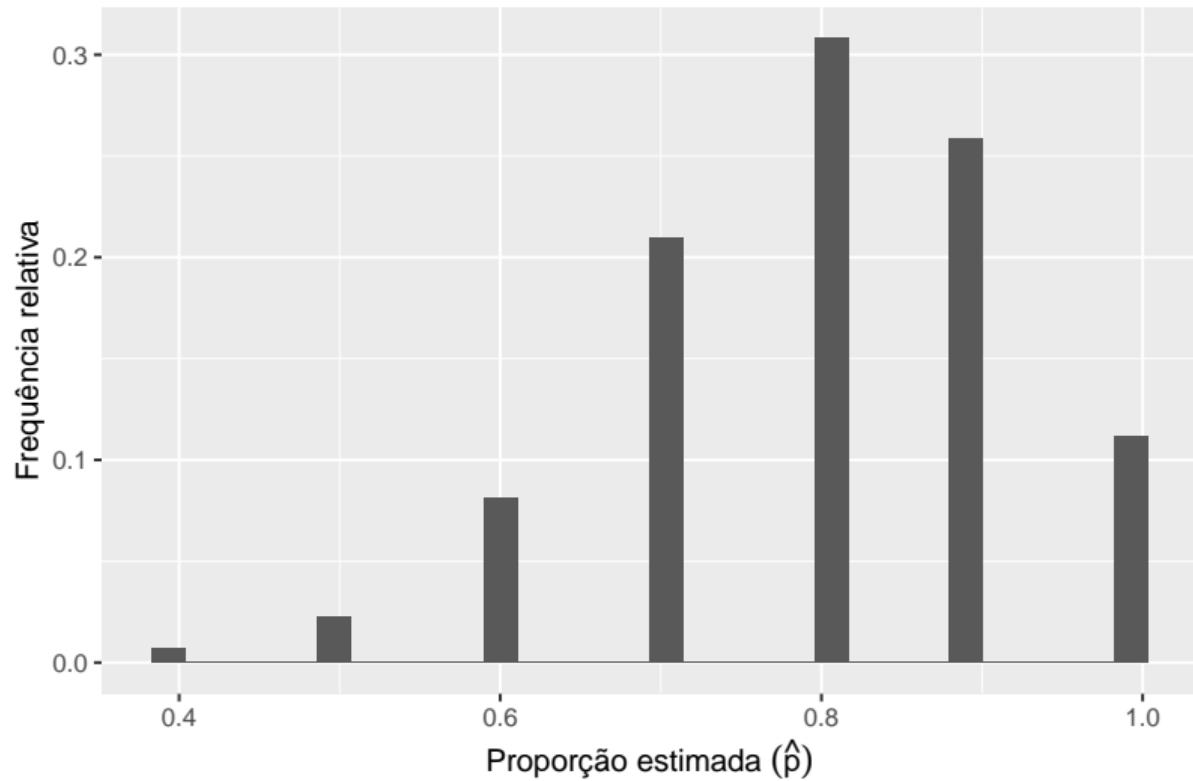


Figura 9. Distribuição amostral da proporção.

# Distribuição amostral

- ▶ Veja que mesmo se o valor verdadeiro for  $p = 0,8$  existe uma probabilidade não desprezível de observarmos 7 pessoas com anticorpos entre as 10 avaliadas.
- ▶ A incerteza associada ao valor de  $p$  no caso de apenas 10 observações é grande.
- ▶ Como podemos diminuir esta incerteza?

# Distribuição amostral

- ▶ Veja que mesmo se o valor verdadeiro for  $p = 0,8$  existe uma probabilidade não desprezível de observarmos 7 pessoas com anticorpos entre as 10 avaliadas.
- ▶ A incerteza associada ao valor de  $p$  no caso de apenas 10 observações é grande.
- ▶ Como podemos diminuir esta incerteza?
- ▶ Solução: Aumentar o número de observações.

# Ilustração computacional

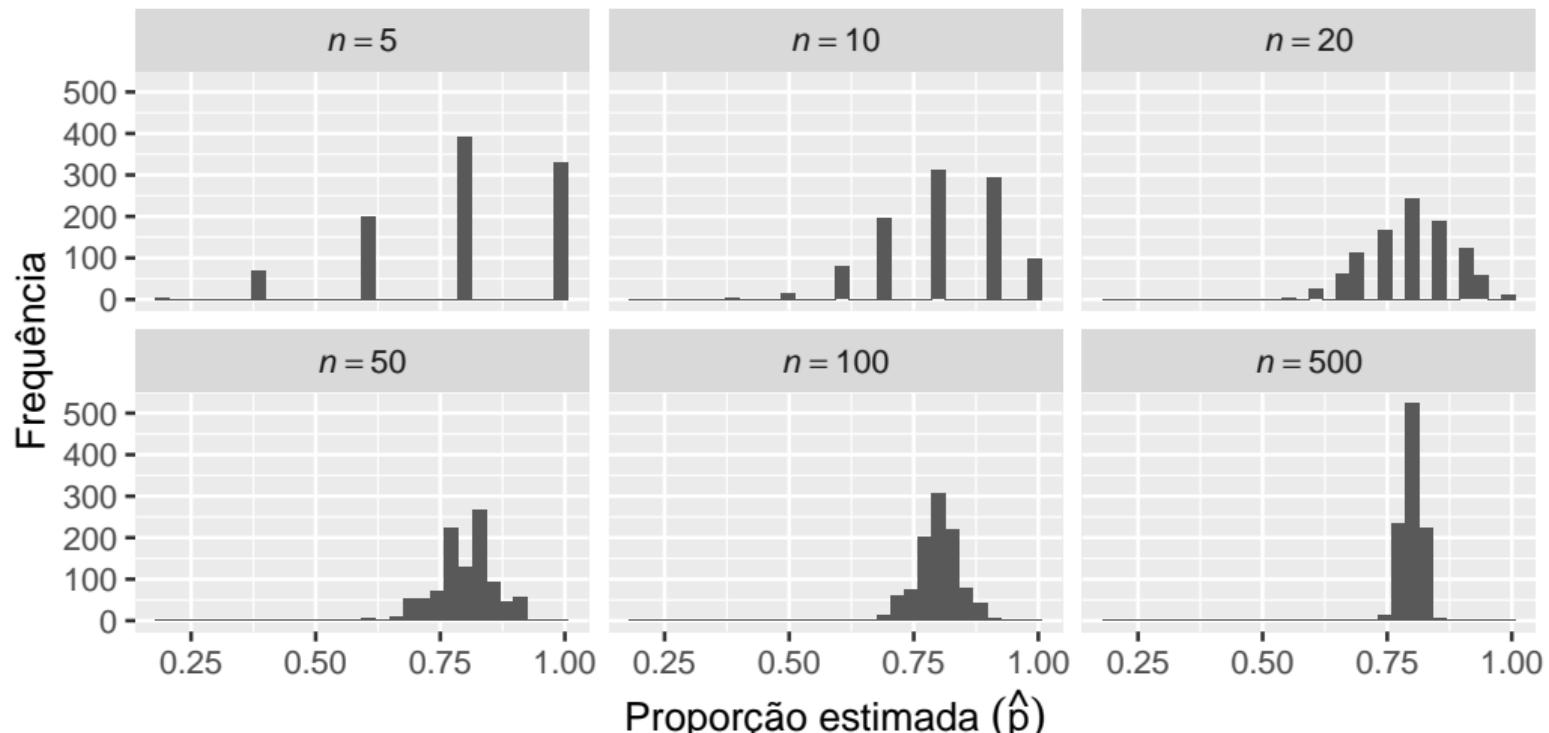


Figura 10. Efeito de aumentar o tamanho da amostra na distribuição amostral da proporção estimada.

# Estatística frequentista

- ▶ Temos o procedimento, mas e como faremos as replicações do experimento em termos práticos?

# Estatística frequentista

- ▶ Temos o procedimento, mas e como faremos as replicações do experimento em termos práticos?
- ▶ Não faremos!!
- ▶ Estimador é função da variável aleatória.
- ▶ Portanto, tem distribuição de probabilidade.
- ▶ A **distribuição amostral** do estimador pode ser usada para estudar o que aconteceria caso o estudo fosse replicado um número muito grande de vezes.
- ▶ Distribuição exata de um estimador é difícil de se obter.
- ▶ O Teorema Central do Limite oferece uma aproximação para amostras grandes (assintótica).

# Reforçando os conceitos

- ▶ Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- ▶ Precisamos saber como a altura dos alunos se distribui.

# Reforçando os conceitos

- ▶ Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- ▶ Precisamos saber como a altura dos alunos se distribui.
- ▶ Formalizando o problema.
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?

# Reforçando os conceitos

- ▶ Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- ▶ Precisamos saber como a altura dos alunos se distribui.
- ▶ Formalizando o problema.
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?
  - ▶  $Y \in \mathbb{R}_+$  - Altura dos alunos da UFPR.
- ▶ Qual a distribuição de probabilidade adequada para esta v.a.?

# Reforçando os conceitos

- ▶ Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- ▶ Precisamos saber como a altura dos alunos se distribui.
- ▶ Formalizando o problema.
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?
  - ▶  $Y \in \mathbb{R}_+$  - Altura dos alunos da UFPR.
- ▶ Qual a distribuição de probabilidade adequada para esta v.a.?
  - ▶ Normal, Gama, Lognormal, Normal Inversa, Weibul, etc.
- ▶ Qual o parâmetro de interesse e o que ele significa?

# Reforçando os conceitos

- ▶ Problema prático: Qual o tamanho ideal de carteiras escolares para os alunos da UFPR?
- ▶ Precisamos saber como a altura dos alunos se distribui.
- ▶ Formalizando o problema.
  - ▶ Qual é a variável aleatória e quais valores ela pode assumir?
  - ▶  $Y \in \mathbb{R}_+$  - Altura dos alunos da UFPR.
- ▶ Qual a distribuição de probabilidade adequada para esta v.a.?
  - ▶ Normal, Gama, Lognormal, Normal Inversa, Weibul, etc.
- ▶ Qual o parâmetro de interesse e o que ele significa?
  - ▶ Altura média e variabilidade da altura dos alunos da UFPR.

# Juntando dados e probabilidades

- ▶ Suponha que uma amostra (AAS) de tamanho  $n = 293$  foi obtida.

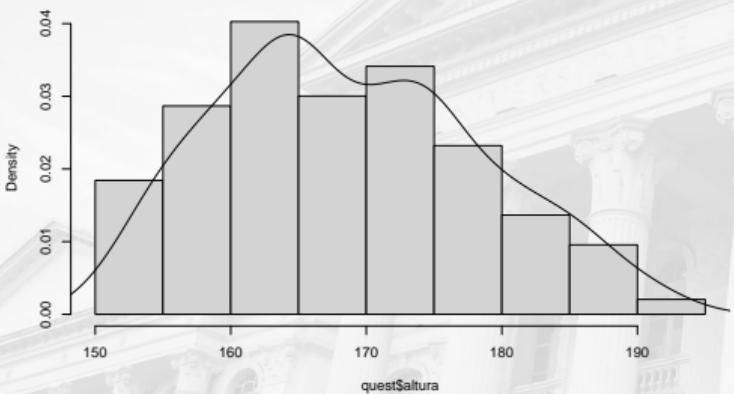


Figura 11. Histograma da altura dos alunos UFPR.

- ▶ Qual modelo é o mais provável de ter gerado essa amostra?

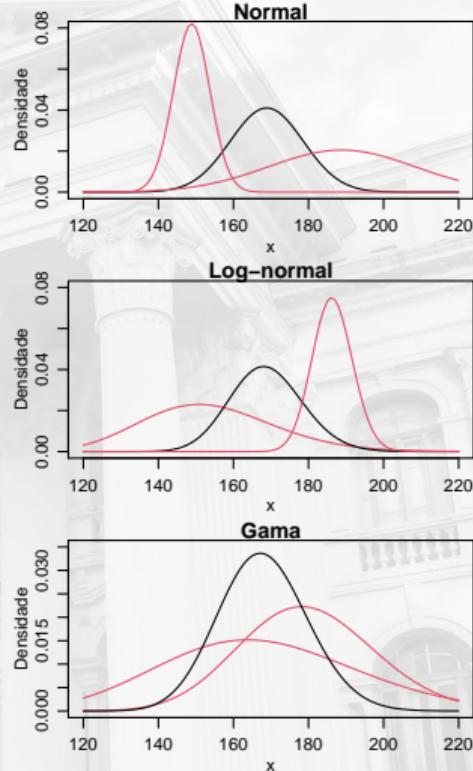


Figura 12. Distribuições de probabilidades candidatas.

# Escolhendo o modelo

- ▶ Modelo Normal
  - ▶ Notação  $Y \sim N(\mu, \sigma^2)$ ;
  - ▶  $f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$ ;
  - ▶  $E(Y) = \mu$  e  $V(Y) = \sigma^2$ .
- ▶ Quais os valores de  $\mu$  e  $\sigma^2$  devo usar?
- ▶ Podemos usar os equivalentes amostrais?
  - ▶  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$  e  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu})^2$ .
- ▶ Como medir a incerteza em  $\hat{\mu}$  e  $\hat{\sigma}^2$ , sendo que temos apenas uma amostra? → Distribuição amostral.
- ▶ E para os outros modelos? → Métodos para estimativação de parâmetros.

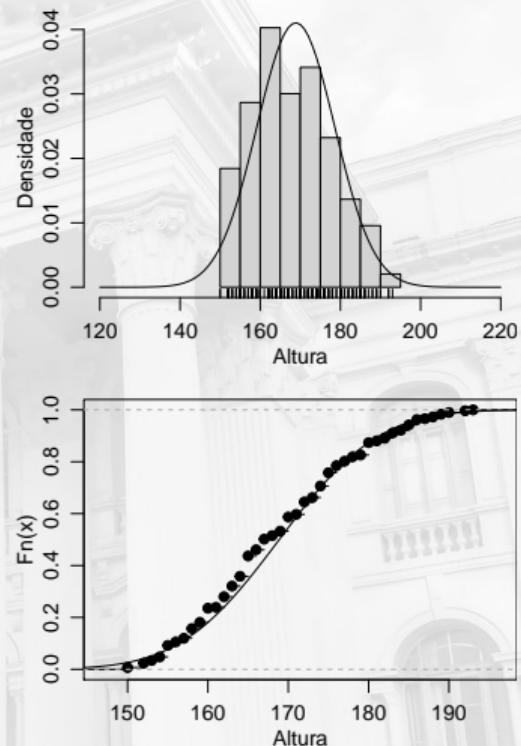


Figura 13. Ajuste da distribuição Normal para a variável altura.

# Distribuição amostral

- ▶ Objeto de inferência (*frequentista*).
- ▶ A estimativa pontual é um resumo desta distribuição.
- ▶ Intervalos entre quantis representam a incerteza sobre o valor estimado.
- ▶ Compara-se estimadores concorrentes pelas características de suas distribuições amostrais.
- ▶ E para tudo isto:  
é preciso saber como estimar.



Figura 14. Distribuição amostral de diferentes estimadores de um parâmetro.

# Resumo

- ▶ Modelo → comportamento da natureza.
- ▶ Parâmetros do modelo → parâmetros populacionais de interesse.
- ▶ Qual modelo melhor descreve os dados?
- ▶ Assumimos um modelo → parâmetros são desconhecidos.
- ▶ Baseado na amostra → encontrar os parâmetros compatíveis com a amostra.
- ▶ Descrever a incerteza → **distribuição amostral**.

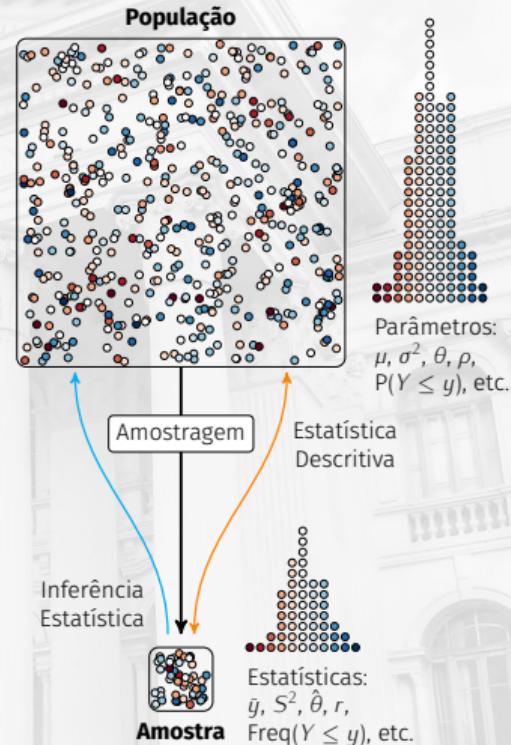


Figura 15. Processo de inferência estatística.